# PRÁCTICA SPARK


**Autores: Mar Lafuente, Elena de la Fuente, Nicolás Machín**
**Mayo 2023**

Universidad Complutense de Madrid
Facultad de Matemáticas
Curso Programación Paralela 2022/2023

**ÍNDICE**

1. Motivación
2. Material utilizado (datos y métodos)
3. Desarrollo
4. Resultados

## 1. MOTIVACIÓN

Esta práctica consiste en el planteamiento, diseño e implementación de una solución a un problema de análisis de datos utilizando Spark, que es un framework de almacenamiento, procesamiento y análisis de datos a gran escala.

El dataset sobre el que trabajaremos será el que proporciona el Ayuntamiento de Madrid del uso del sistema de bicicletas de préstamo BICIMAD. En concreto, nos centraremos en el año 2019, incluyendo los 12 meses.

A continuación, veremos por ejemplo cuales son los grupos (según el rango de edad) que más utilizan el servicio de Bicimad, cual es la media de tiempo de uso en total y según cada grupo, cual ha sido la estación más concurrida etc.

Nuestro trabajo se dividirá en dos archivos distintos: ***EstudioGeneral.py*** y ***EstudioGrupoEdad.py***. Los dos archivos tienen una cierta similitud a excepción de algunas funciones. El primero consistirá en un análisis centrado en función de las estaciones y meses del año y el segundo en función de los resultados anuales.

## 2. MATERIAL UTILIZADO (DATOS Y MÉTODOS)

Librerías: el código está implementado en el lenguaje Python y haremos uso de:

- *pyspark.sql* : biblioteca de Python para usar *Spark*
- *matplotlib.pyplot* : librería para la representación de gráficas en Python

Archivos: de la web Open Data de la EMT Madrid, trabajaremos con el dataset de 2019

1. *BiciMAD_movements_2019_7_12/201901_Usage_Bicimad.json*
2. *BiciMAD_movements_2019_7_12/201902_Usage_Bicimad.json*
3. *BiciMAD_movements_2019_7_12/201903_Usage_Bicimad.json*
4. *BiciMAD_movements_2019_7_12/201904_Usage_Bicimad.json*
5. *BiciMAD_movements_2019_7_12/201905_Usage_Bicimad.json*
6. *BiciMAD_movements_2019_7_12/201906_Usage_Bicimad.json*
7. *BiciMAD_movements_2019_7_12/201907_movements.json*
8. *BiciMAD_movements_2019_7_12/201908_movements.json*
9. *BiciMAD_movements_2019_7_12/201909_movements.json*
10. *BiciMAD_movements_2019_7_12/201910_movements.json*
11. *BiciMAD_movements_2019_7_12/201911_movements.json*
12. *BiciMAD_movements_2019_7_12/201912_movements.json*

Interpretación de los archivos de datos de Bicimad:

**_id** : Identificador del movimiento.

**user_day_code** : Código del usuario. Para una misma fecha, todos los movimientos de un mismo usuario tendrán el mismo código.

**idunplug_station** : Número de la estación de la que se desengancha la bicicleta.

**idunplug_base** : Número de la base o enganche de la que se desengancha la bicicleta.

**idplug_station** : Número de la estación en la que se engancha la bicicleta.

**idplug_base** : Número de la base o enganche en la que se engancha la bicicleta.

**unplug_hourTime** : Franja horaria en la que se realiza el desenganche de la bicicleta. Por cuestiones de anonimato, se facilita la hora de inicio del movimiento, sin la información de minutos y segundos. Todos los movimientos iniciados durante la misma hora tendrán el mismo dato de inicio.

**travel_time** : Tiempo total en segundos, entre el desenganche y el enganche de la bicicleta.

**user_type** : Número que indica el tipo de usuario que ha realizado el movimiento. Sus posibles valores son:
- 0: No se ha podido determinar el tipo de usuario
- 1: Usuario anual
- 2: Usuario ocasional
- 3: Trabajador de la empresa

**ageRange** : Número que indica el rango de edad del usuario que ha realizado el movimiento.
- 0: No se ha podido determinar el rango de edad del usuario
- 1: El usuario tiene entre 0 y 16 años
- 2: El usuario tiene entre 17 y 18 años
- 3: El usuario tiene entre 19 y 26 años
- 4: El usuario tiene entre 27 y 40 años
- 5: El usuario tiene entre 41 y 65 años
- 6: El usuario tiene 66 años o más

**zip_code** : código postal del barrio al que pertenece la estación de salida.

## 3. DESARROLLO

Funciones definidas en los dos scripts:

- **columnas** : dado un dataframe, añade una columna mes y estación del año, dependiendo de la naturalidad del archivo y selecciona las columnas relevantes.

- **union** : dada una lista de dataframes, los encadena y une todos en uno solo.

- **resultados_descriptivo** : función que muestra por pantalla los resultados obtenidos.

- **grafico_barras, grafico_sectores** y **mostrar_grafica** : funciones definidas para la representación de gráficas (columnas y circular) de diferentes variables y datos.

- **buscar_media_total** : dado un dataframe y una variable, devuelve la media de la variable en cuestión.

Funciones definidas en ***EstudioGeneral.py***:

- **count_var** : dado un dataframe y una variable, devuelve un dataframe con la cantidad de veces referente a la variable, el valor máximo y su grupo correspondiente, el valor mínimo y su grupo correspondiente

- **porcent_count_age** y **porcent_count_age1** : funciones que devuelven el porcentaje de ageRange en función de la época del año.

- **df_show** : dada una lista de porcentajes, muestra por pantalla un dataframe

Funciones definidas en ***EstudioGrupoEdad.py***:

- **count_age** : dado un dataframe, devuelve un dataframe con la cantidad de veces que cada ageRange usa BiciMad, el grupo correspondiente al valor máximo de uso y el valor máximo, el grupo correspondiente al valor mínimo de uso y su valor mínimo.

- **porcent_count_age** : dado un dataframe, devuelve una lista con el porcentaje de veces que cada ageRange usa el servicio de Bicimad en una lista ordenada.

- **estudio_porcent** : dado un dataframe y una función f, devuelve la tabla con los porcentajes y el gráfico de sectores con esos porcentajes.

- **buscar_max_age** : dado un dataframe devuelve el dataframe con los tiempos máximos de uso cada grupo de edad y el grupo de edad que haya hecho el viaje más largo.

- **buscar_mediaVar_age** : dado un dataframe y una variable (en general un string, correspondiente al título de la columna del dataframe), devuelve un dataframe con las medias de la variable para cada grupo de edad. Devuelve también el valor máximo y mínimo con su grupo correspondiente.

- **porcent_age_time** : dado un dataframe y una variable, devuelve una lista con el porcentaje de tiempo que cada ageRange usa Bicimad en una lista ordenada.

- **station_min_max** y **station_min_max1** : dado un dataframe y una variable (correspondiente a 'idplug_station' o 'idunplug_station'), devuelve la estación más y menos concurrida de cada una de las dos variables.

## 4. RESULTADOS

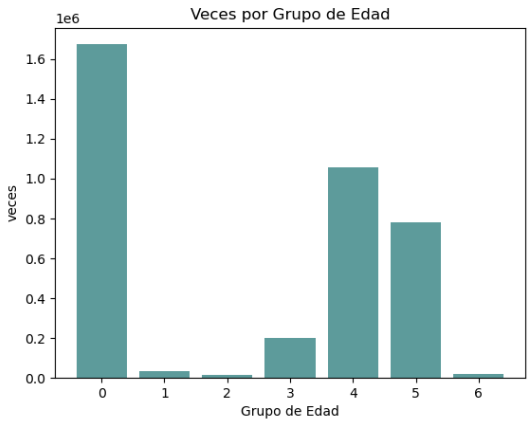
Una vez ejecutados los scripts, obtenemos los siguientes resultados:

***EstudioGrupoEdad***

### Rango de Edad

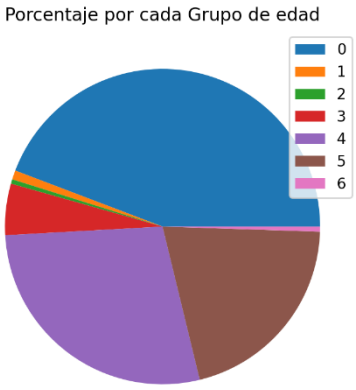
Tabla de **frecuencias** de usuarios dividido en Grupo de Edad del año 2019:

```
+--------+-------+
|ageRange|  count|
+--------+-------+
|       0|1674369|
|       1|  35706|
|       2|  17190|
|       3| 199364|
|       4|1057029|
|       5| 783090|
|       6|  19627|
+--------+-------+
```




El grupo de Edad con mayor uso en 2019 ha sido el 0 con un total de 1674369 veces.
El grupo de Edad con menor uso en 2019 ha sido el 2 con un total de 17190 veces.

| ageRange | count percentage |
|---|---|
| 0 | 44.22 |
| 1 | 0.94 |
| 2 | 0.45 |
| 3 | 5.27 |
| 4 | 27.92 |
| 5 | 20.68 |
| 6 | 0.52 |




**Interpretación:** observamos el grupo que más utiliza 'Bicimad' es el 0, lo cual nos indica que no se ha podido determinar la edad del usuario. Esto significa que la aplicación de 'Bicimad' necesita una mejoría en cuanto a ese campo.
Sin embargo, en las gráficas se puede ver que el siguiente grupo con mayor frecuencia es el correspondiente a las personas de 27 a 40 años (grupo 4); en cambio los que menos valores han obtenido son los de 17 y 18 años (grupo 2). Por lo tanto, podemos asumir que estos resultados tienen una estrecha relación con la diferencia del rango de edad escogidos para cada grupo.

Tabla de **tiempos máximos** recorridos dividido en Grupo de Edad del año 2019:

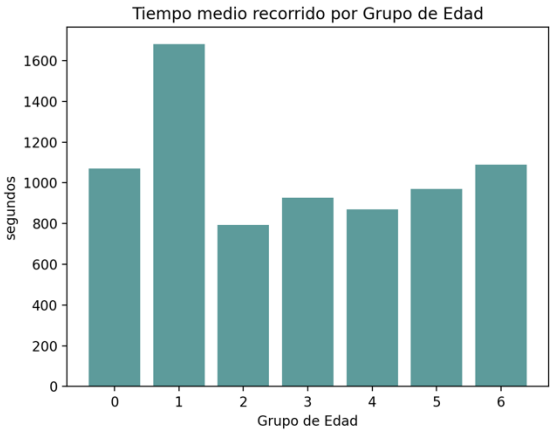
| ageRange | max(ageRange) | max(travel_time) |
|---|---|---|
| 0 | 0 | 86291 |
| 1 | 1 | 85151 |
| 2 | 2 | 21473 |
| 3 | 3 | 74350 |
| 4 | 4 | 85699 |
| 5 | 5 | 86399 |
| 6 | 6 | 20971 |

Grupo de Edad con mayor tiempo en 2019 ha sido 5 con un total de 86399 segundos (1 día aprox.). La media de tiempo de uso respecto al total de usuarios es:  990.28 segundos (16 minutos aprox.).

**Observación:** Con los datos anteriores es fácil ver que pese a acotar superiormente el tiempo de recorrido a menos de un día, hemos obtenido resultados iguales a esa cota; sin embargo, la media se encuentra en 16 minutos. Esto significa que existen observaciones atípicas, por lo que proseguiremos el estudio con valores promedio, puesto que son más significativos y fiables.

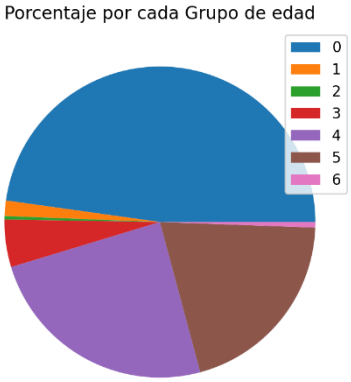
Tabla de **tiempo medio recorrido** por los usuarios dividido en Grupo de Edad del año 2019:

| ageRange | avg(travel_time) |
|---|---|
| 0 | 1070.0904448183167 |
| 1 | 1680.20741612054 |
| 2 | 793.9428737638161 |
| 3 | 927.5768694448345 |
| 4 | 869.0238451357532 |
| 5 | 969.6169712293606 |
| 6 | 1089.305548479136 |




Grupo de Edad con mayor tiempo medio en 2019 ha sido 1 con un total de casi 1680 segundos (28 minutos aprox.).
Grupo de Edad con menor tiempo medio en 2019 ha sido 2 con un total de casi 794 segundos (13 minutos aprox.).

| ageRange | count percentage |
|---|---|
| 0 | 47.785053 |
| 1 | 1.6000167 |
| 2 | 0.36398673 |
| 3 | 4.931932 |
| 4 | 24.498474 |
| 5 | 20.25034 |
| 6 | 0.57019585 |




**Interpretación:** vemos en la gráfica que el grupo 1 (de 0 a 16 años) destaca en cuanto a recorridos más largos en promedio. Esto puede deberse al no disponer de carnet de conducir de ningún tipo, ya que puede ser una alternativa más económica y accesible para poder circular por la ciudad.

## Estaciones Bicimad

Finalmente, en el script hemos obtenido los siguientes resultados en cuanto a la concurrencia de estaciones en Madrid:
La estación de llegada menos usada es la número 2008 y la más concurrida es 175.
La estación de salida menos usada es la número 2008 y la más concurrida es 175.

**Interpretación:** podemos suponer que la estación más frecuentada deber estar situada en una zona turística en Madrid, al igual que podemos pensar que la menos concurrida estará en una zona menos poblada. Esto se debe que han coincidido los resultados tanto en la variable de llegada como en la de salida. Por otro lado, es posible

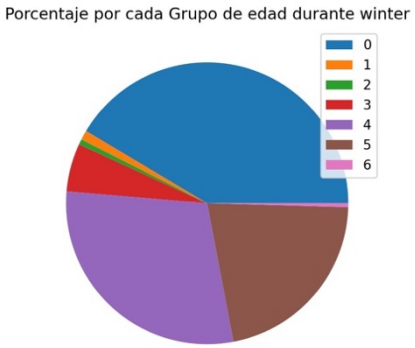
que la capacidad de bicicletas de cada estación también sea un influyente en la concurrencia.

***EstudioGeneral.py***

## Estacionalidad

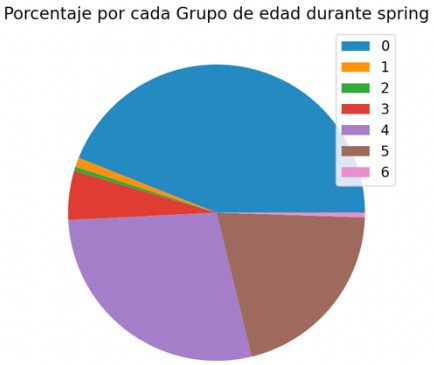
Los porcentajes en función del count y del ageRange según la estación invierno es:

| ageRange | count percentage |
|---|---|
| 0 | 41.43 |
| 1 | 1.1 |
| 2 | 0.61 |
| 3 | 5.51 |
| 4 | 29.35 |
| 5 | 21.57 |
| 6 | 0.44 |



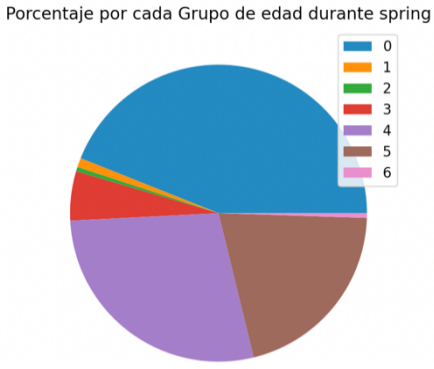

Los porcentajes en función del count y del ageRange según la estación primavera es:

| ageRange | count percentage |
|---|---|
| 0 | 43.96 |
| 1 | 1.01 |
| 2 | 0.47 |
| 3 | 5.36 |
| 4 | 28.01 |
| 5 | 20.67 |
| 6 | 0.51 |



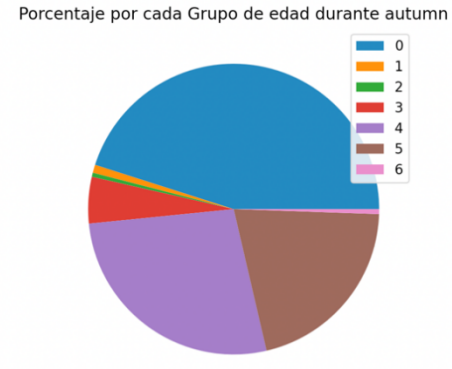

Los porcentajes en función del count y del ageRange según la estación verano es:

| ageRange | count percentage |
|---|---|
| 0 | 44.72 |
| 1 | 0.86 |
| 2 | 0.36 |
| 3 | 5.17 |
| 4 | 28.39 |
| 5 | 19.97 |
| 6 | 0.52 |



Los porcentajes en función del count y del ageRange según la estación otoño es:

| ageRange | count percentage |
|---|---|
| 0 | 45.08 |
| 1 | 0.9 |
| 2 | 0.45 |
| 3 | 5.17 |
| 4 | 27.04 |
| 5 | 20.82 |
| 6 | 0.55 |

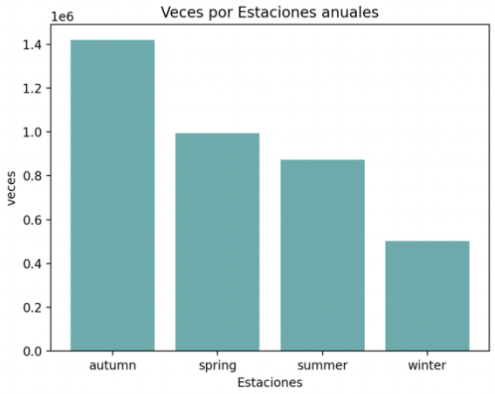



La media de tiempo de un viaje durante invierno es 1112 segundos o 18.52 minutos
La media de tiempo de un viaje durante primavera es 1153 segundos o 19.21 minutos
La media de tiempo de un viaje durante verano es 1154 segundos o 19.22 minutos
La media de tiempo de un viaje durante otoño es 1160 segundos o 19.33 minutos

**Interpretación:** hemos comprobado que el estudio en las distintas estaciones no supone una diferencia significativa en los anteriores porcentajes. Es por ello que, si se quisiera profundizar más en el estudio de estos, recomendaríamos hacerlo de manera anual.

Tabla de frecuencias de usuarios dividido en Estaciones anuales del año 2019:

| season | count |
|---|---|
| autumn | 1418467 |
| spring | 994221 |
| summer | 872388 |
| winter | 502608 |



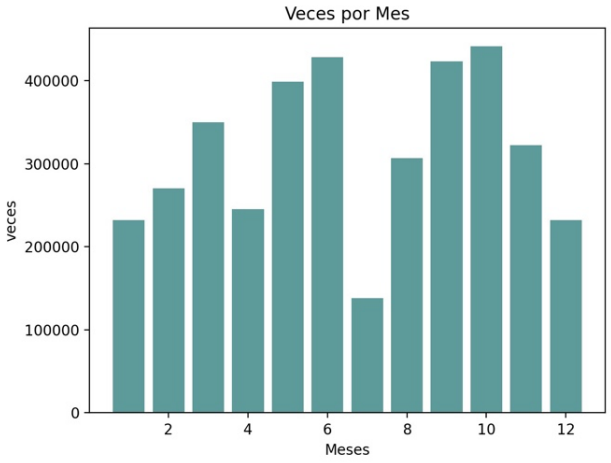

La estación con mayor uso en 2019 ha sido otoño con un total de 1418467 veces.
La estación con menor uso en 2019 ha sido invierno con un total de 502608 veces.

**Interpretación:** constatamos en la gráfica que las estaciones primavera y otoño sobresalen en comparación al resto. Podría estar relacionado con la diferencia de temperaturas de la ciudad alcanzadas durante las estaciones de verano e invierno.

## MESES

Tabla de frecuencias de usuarios dividido en Meses del año 2019:

| month | count |
|---|---|
| 1 | 232008 |
| 2 | 270600 |
| 3 | 349979 |
| 4 | 245527 |
| 5 | 398715 |
| 6 | 427946 |
| 7 | 137854 |
| 8 | 306588 |
| 9 | 422969 |
| 10 | 441107 |
| 11 | 322431 |
| 12 | 231960 |




Meses con mayor uso en 2019 ha sido 10 con un total de 441107 veces.
Meses con menor uso en 2019 ha sido 7 con un total de 137854 veces.

**Interpretación:** visualizamos en los datos que septiembre es el mes con mayor cantidad de usuarios y julio el que menos. Esto se puede deber al regreso de las vacaciones o del comienzo de la actividad académica y la usual intención de lograr nuevos propósitos a nivel personal, tal como ser más ecológico y obtener una rutina de ejercicio. No obstante, de nuevo creemos que el clima puede un considerable influyente de los resultados, puesto que por el calor de julio no es recomendable el deporte en un mayor rango de horas del día, pero en septiembre tiempo es más favorable y continúa el turismo de verano.

## 5. Conclusiones

1. Debido a que en la gran mayoría de datos se desconoce la edad del usuario, recomendamos una mejora en la obtención de ese campo para un futuro estudio que tenga en cuenta la variable ‘ageRange’.
2. Además, creemos que los rangos de edad deben estar más equilibrados y mejor definidos unos con otros para que sean un mejor reflejo de la población.
3. Finalmente, pensamos que la temperatura y el turismo son factores importantes en cuanto a la cantidad de usuarios con respecto a la utilización del servicio de Bicimad. Es por ello que se debería tener en cuenta la época del año a la hora de realizar posteriores estudios.